SANDA CHERATA, MANUELA MIHĂESCU

# MODELE FORMALE DE REPREZENTARE A INFORMAȚIILOR LEXICALE ȘI TERMINOLOGICE ÎN PROIECTUL *SIASTRO*

## 1. Introducere

În momentul de față există o varietate largă de aplicații foarte complexe de tratare, cu mijloace informatice, a limbajului natural – NLP [Natural Language Processing]. Așa cum se arată în G. Francopoulo *et alii* (2006a), optimizarea producerii, întreținerii și extinderii resurselor lexicale este unul dintre aspectele cruciale ale aplicațiilor NLP. Un al doilea aspect implică optimizarea procesului de integrare a acestor resurse în aplicațiile lingvistice. De-a lungul timpului s-au produs, cu eforturi mari, numeroase resurse lexicale pentru diverse limbi. Ar fi benefic (Segura Bedmar *et alii*, 2006) ca toate aceste resurse să poată fi integrate pentru a forma resurse globale extinse.

Pe de altă parte, dezvoltarea tehnico-științifică a dus la apariția terminologiilor specifice diverselor domenii de specialitate, iar procesul de globalizare a generat, pe lângă schimbul de produse și tehnologii, și schimbul de documentații. În acest context, aplicațiile de gestiune terminologică au o importanță crucială.

Până în ultimul deceniu, aplicațiile NLP și cele terminologice au urmat un curs paralel, fiecare reprezentându-și structurile de date de așa manieră încât să răspundă cerințelor formale și condițiilor tehnice existente în momentul respectiv. Informațiile din lexicoanele aplicațiilor NLP vizează în primul rând atributele gramaticale (lexicale, morfologice, sintactice și semantice) și, foarte puțin, atribute specifice termenilor. Aplicațiile de gestiune terminologică (marea lor majoritate abordând orientarea conceptuală în terminologie) uzează foarte puțin de informația gramaticală, prezentând pe larg, în schimb, atribute ale termenilor: domeniul, sistemul conceptual, conceptul desemnat, definiție, contexte, informații de utilizare, statut, autoritate normativă etc.

Evoluția tehnică și științifică a dus la apariția unor sisteme informatice din ce în ce mai complexe. S-a simțit nevoia ca, pe de o parte, aplicațiile NLP să poată exploata termenii (care necesită o tratare specială, fiind, în general, expresii multi-cuvânt care funcționează ca unități de sine stătătoare), iar pe de altă parte, aplicațiile de gestiune terminologică să poată utiliza facilitățile oferite de aplicațiile

*DACOROMANIA, serie nouă, XIII, 2008, nr. 2, Cluj-Napoca, p. 151–169*

NLP (unele dintre cele mai utile fiind recunoaşterea terminologică avansată şi extragerea termenilor). Mai mult decât atât, în prezent se caută metode de a facilita interschimbul de date între diverse produse informatice. S-a ajuns astfel să se caute elaborarea unor modele de reprezentare a informaţiilor lexicale care să permită exploatarea acestora de către cele mai variate aplicaţii, atât de natură terminologică, cât şi din clasa NLP. Există mai multe modele de organizare a informaţiilor care încearcă să răspundă acestor exigenţe, unele dintre ele (de exemplu LMF[1]) în curs de standardizare ISO. Fiecare are avantaje şi dezavantaje, iar cercetările în domeniu sunt foarte dinamice.

Proiectul **SIASTRO**[2], care are ca temă realizarea unui sistem informatic pentru analiza sintagmatică a textelor în limba română, îşi propune, ca o primă aplicaţie practică, implementarea unui extractor de termeni şi înregistrarea acestora într-o colecţie de date terminologice. Prin urmare, el se află la punctul de întâlnire a NLP cu sistemele de gestiune terminologică. S-a căutat să se abordeze un mod de structurare şi reprezentare a datelor lingvistice şi terminologice în conformitate cu modelele formale existente. Acest lucru permite extinderea şi, eventual, integrarea aplicaţiilor noastre în cele dezvoltate în acest domeniu.

Am început, aşadar, demersul nostru prin analiza modelelor existente, căutând să alegem un mod de reprezentare care să fie cât mai apropiat unui model standard, să răspundă cerinţelor analizei sintagmatice pentru limba română şi, totodată, să adapteze şi să extindă atributele *Dicţionarului Morfologic Român*[3] realizat de colectivul RoLingva[4].

### 1.1. Standarde şi modele pentru resursele lexicale

Există numeroase încercări de a standardiza procesele şi resursele lingvistice. Amintim aici câteva dintre cele mai importante proiecte care au avut drept scop crearea de modele pentru reprezentarea resurselor lingvistice:

- GENELEX (1990-1994) – a fost un proiect EUREKA, unul dintre scopurile sale fiind acela de a proiecta un model global de reprezentare a oricărui tip de informaţie lexicală, într-un mod neutru, independent de aplicaţie şi neataşat unei teorii lingvistice specifice.
- EAGLES[5] (Expert Advisory Group on Language Engineering Standards) – o iniţiativă a Comisiei Europene în cadrul programului DG XIII *Linguistic*

[1] Lexical Mark-up framework: http://tagmatica.fr./doc/ISO24613cdRev9.pdf

[2] Proiectul *Sistem informatic pentru analiza sintagmatică a textelor în limba română. Fundamentare teoretică şi implementare – SIASTRO* se desfăşoară în cadrul programului *86 CEEX-II 03 / 31.07.2006.* Realizatorii sunt: Universitatea „Babeş-Bolyai” din Cluj-Napoca (Facultatea de Litere), coordonator, Software ITC SA (prin colectivul *RoLingva*), Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu” din Cluj-Napoca şi Universitatea Tehnică din Cluj-Napoca (Catedra de Calculatoare). Vezi *supra*, Tămâianu-Morita 2008.

[3] http://www.rolingva.ro/aplicatii_dictionar.php

[4] http://www.rolingva.ro/

[5] http://www.ilc.cnr.it/EAGLES/home.html

*Research and Engineering*, a avut ca scop accelerarea producerii de standarde pentru:
  - o resurse lingvistice la scară foarte largă;
  - o mijloacele de tratare a unor astfel de resurse prin formalisme ale lingvisticii computaţionale şi aplicaţii informatice diverse;
  - o mijloace de atestare şi evaluare a unor astfel de resurse, aplicaţii şi produse.
- ISLE[6] (International Standards for Language Engineering) – este atât numele unui proiect, cât şi numele unui întreg set de activităţi coordonate din domeniul tehnologiei limbajului uman (HLT – Human Language Technology). ISLE a funcţionat sub egida EAGLES; scopul său a fost să dezvolte standarde într-un cadru internaţional, în contextul iniţiativei EU-US International Research Cooperative.
- MULTEX[7] (Multilingual Text Tools and Corpora) – înglobează mai multe proiecte care vizează elaborarea de standarde pentru codificarea textelor şi a corpusurilor şi dezvoltarea aplicaţiilor şi a resurselor lingvistice; proiectul a stabilit linii directoare pentru NLP şi traducerea automată (MT – Machine Translation) bazată pe corpusuri.
- PAROLE (Preparatory Action for Linguistic Resources Organisation for Language Engineering) – proiect al Uniunii Europene, a avut ca scop armonizarea resurselor lexicale şi a corpusurilor pentru limbile Uniunii Europene.
- SIMPLE[8] – proiect sponsorizat de Programul Cadru IV, a reprezentat prima încercare de a dezvolta lexicoane semantice pentru un număr mare de limbi, cu un model comun care codifică „tipuri semantice” structurate şi cadre semantice; dezvoltat în strânsă legătură cu PAROLE.

### 1.2. Standarde şi modele pentru resursele terminologice

Marea diversitate a reprezentării informaţiilor terminologice a dus la apariţia unor standarde internaţionale care stabilesc principii, metode de lucru şi modele de reprezentare pentru a se putea realiza diseminarea şi interschimbul datelor terminologice. În cadrul ISO (*International Organisation for Standardisation*) a fost creat comitetul tehnic 37 (TC37 – *Terminology and other Language Resources*) care a elaborat numeroase standarde, dintre care amintim:
- ISO 704:2000 *Terminology work – Principles and methods;*
- ISO 12620:1999 *Computer applications in terminology – Data categories;*
- ISO 1087-1:2000 *Terminology work – Vocabulary – Part 1: Theory and application;*

[6] http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
[7] http://aune.lpl.univ-aix.fr/projects/multext/
[8] http://www.ub.es/gilcub/SIMPLE/simple.html

- ISO 1087-2:2000 *Terminology work – Vocabulary – Part 2: Computer applications;*
- ISO 12616:2002 *Translation-oriented terminography;*
- ISO 12200:1999 *Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange;*
- ISO/FDIS 24613 *Language resource management–Lexical markup framework(LMF)*.
  www.tc37sc4.org/new_doc/ISO_TC37_N130_rev9_LMF_15March2006.pdf

1.3 Modele pentru integrarea resurselor lexicale şi a celor terminologice

Există, de asemenea, o mulţime de iniţiative a diverselor organizaţii şi consorţii, care îşi propun să atingă acelaşi scop: reprezentarea uniformă a informaţiilor lexicale în vederea realizării interschimbului între diverse aplicaţii şi a reutilizării resurselor, precum şi integrarea informaţiilor terminologice şi a celor lexicale în aceeaşi gamă de aplicaţii.

**TEI**[9] (Text Encoding Initiative), lansat iniţial în 1987, este un standard internaţional şi interdisciplinar de reprezentare a informaţiilor textuale. Din 2000 s-a înfiinţat un consorţiu care-i sprijină activitatea. Standardul TEI prevede modele de reprezentare atât a informaţiilor lexicale, cât şi a celor terminologice.

Consorţiul **SALT**[10] (Standards-based Access service to multilingual Lexicons and Terminologies), format din grupuri academice, guvernamentale, comerciale şi asociaţii din Europa şi SUA a fost creat tocmai pentru aceasta şi a elaborat formatul **XLT**.

În acelaşi scop se desfăşoară activitatea grupului **OSCAR**[11] (Open Standards for Container/Content Allowing Re-use) din cadrul asociaţiei **LISA** (Localisation Industry Standards Association).

**OLIF**[12] este un consorţiu sub egida căruia s-a elaborat un standard deschis cu acelaşi nume (OLIF – Open Lexicon Interchange Format), pentru interschimbul de informaţii lexicale şi terminologice.

## 2. Modelele analizate în cadrul proiectului SIASTRO

Dintre modelele de reprezentare a informaţiei lexicale ne-am oprit la studierea **LMF** (Lexical Markup Framework) şi **OLIF** (Open Lexicon Interchange Format**)**, care se apropie cel mai mult atât de cerinţele formale ale gramaticii limbii române, cât şi de cele impuse de realizarea tehnică a obiectivului propus.

---

[9] http://www.tei-c.org/index.xml
[10] http://www.ttt.org/salt/index.html
[11] http://www.lisa.org/sigs/oscar/
[12] http://www.olif.net/

### *2.1. LMF – Lexical Markup Framework*

**LMF** – Lexical Markup Framework este o propunere de standard ISO 24613:2006. În momentul de față este un document de lucru. Lexical Markup Framework este un metamodel abstract care oferă un cadru comun, standardizat, pentru construcția lexicoanelor computerizate. LMF asigură codificarea informației lingvistice într-un mod care permite reutilizarea ei în diverse aplicații. El oferă o reprezentare comună, care poate fi partajată între aplicații, a obiectelor lexicale, incluzând aspectele morfologice, sintactice şi semantice. LMF răspunde unor cerințe de normalizare a lexicoanelor utilizate în aplicațiile NLP (Francopoulo *et alii* 2006, p. 27) astfel:

1. Reprezintă cuvinte în limbi în care sunt posibile mai multe ortografii (nativă sau transliterată).
2. Reprezintă morfologia unor limbi pentru care reprezentarea tuturor formelor flexionale nu poate fi gestionată; pentru astfel de limbi, doar reprezentarea în intensiune este singura de luat în considerare.
3. Asociază uşor formele scrise cu formele rostite pentru orice limbă.
4. Reprezintă cuvinte compuse complexe.
5. Reprezintă expresii multi-cuvânt fixe, semi-fixe sau flexibile.
6. Reprezintă comportări sintactice specifice.
7. Permite corespondența complexă a argumentelor între descrierile sintactice şi semantice.
8. Permite organizarea semantică bazată pe SynSet-uri (ca în WordNet) sau pe predicate semantice (ca în FrameNet).
9. Reprezintă resurse lingvistice multilingve de dimensiuni mari pe baza pivoților inter-limbi sau pe baza legăturilor de transfer.

LMF utilizează Unicode ca sistem de codificare a caracterelor. Constantele lingvistice sunt specificate în Data Category Registry (DCR), aşa cum sunt definite de ISO TC37. Specificația LMF este conformă principiilor de modelare ale **UML** (Unified Modeling Language)[13].

LMF este compus din următoarele componente:

- Un model central (**core package**), care cuprinde un metamodel, adică scheletul structural al LMF; acesta descrie ierarhia informațiilor incluse într-o intrare lexicală. Modelul central este completat cu resurse diverse care fac parte din definiția LMF. Aceste resurse cuprind:
  - categorii de date specifice, utilizate de varietatea tipurilor de resurse asociate cu LMF; aceste categorii de date sunt relevante pentru metamodel şi sunt asociate cu extensii ale modelului central;

[13] UML a fost elaborat în 1997 de *Object Management Group* (OMG) cu scopul de a oferi un limbaj de proiectare comun şi stabil, care să servească la dezvoltarea aplicațiilor informatice. UML a devenit un limbaj de modelare standard.

- restricțiile care guvernează relațiile categoriilor de date cu metamodelul şi cu extensiile sale;
- proceduri standard pentru exprimarea categoriilor de date în XML.

• Extensiile modelului central:
  - extensii pentru lexicoane electronice [en: machine readable lexicons];
  - extensii pentru lexicoanele NLP.

*2.1.1. Metamodelul LMF*

Modelul LMF este format din clase UML, asocieri între clase, precum şi un set de categorii de date conforme cu ISO 12620, reprezentate ca perechi atribut-valoare.

Cei care creează un lexicon trebuie să utilizeze clasele specificate în pachetul central LMF (**LMF core package**). Se pot utiliza clasele suplimentare definite în extensiile standardului (**LMF extensions**). De asemenea, cei care creează un lexicon trebuie să definească selecția categoriilor de date.

**2.1.1.1**. **Pachetul central LMF**

Pachetul central LMF este un metamodel care oferă un mijloc flexibil de a construi modele LMF şi extensii. Reprezentarea UML a pachetului central este următoarea[14]:

Clasa *DataBase* este o clasă unică şi reprezintă resursa lingvistică în întregime. DataBase este container pentru unul sau mai multe lexicoane.

Clasa *Lexicon* este container pentru toate intrările lexicale ale unei limbi în baza de date. Un *Lexicon* trebuie să conțină cel puțin o intrare lexicală (*Lexical Entry*).

Clasa *Lexical Information* conține informații administrative şi alte atribute generale. Această clasă descrie informațiile generale referitoare la un lexicon.

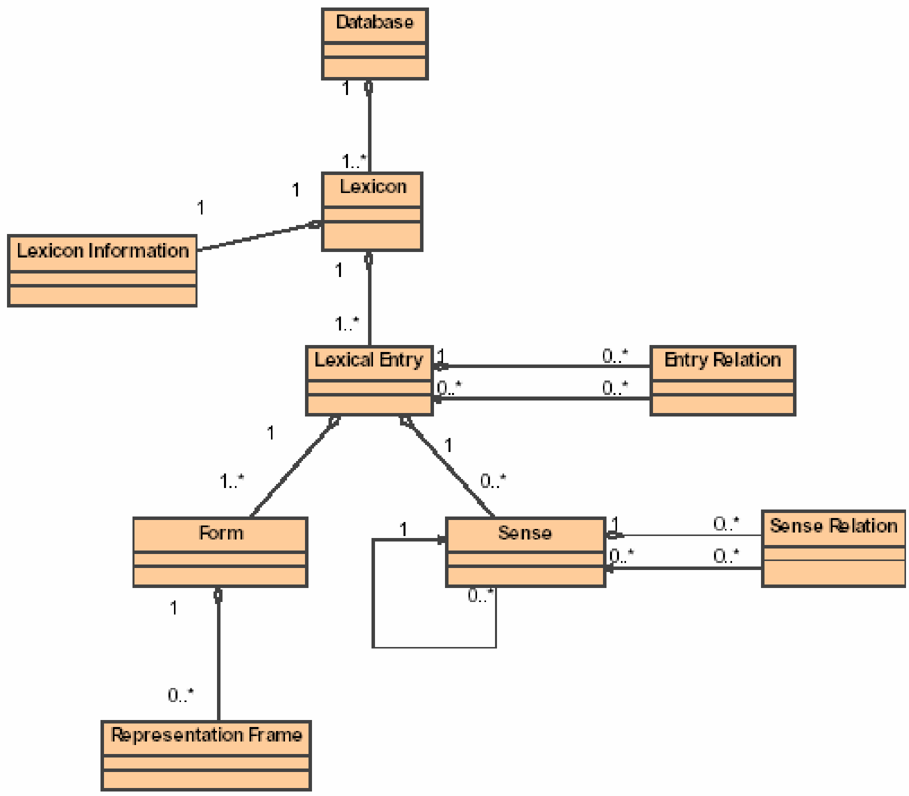
Clasa *Lexical Entry* este un container pentru reprezentarea componentelor de nivel superior ale limbii. În consecință, numărul de reprezentări pentru cuvintele simple, expresiile multi-cuvânt şi afixe din lexicon este egal cu numărul intrărilor lexicale din lexiconul respectiv. *Lexical Entry* este un container pentru tratarea claselor *Form* şi *Sense*. Prin urmare, *Lexical Entry* tratează relațiile dintre forme şi semnificațiile asociate lor. O intrare lexicală – *Lexical Entry* – poate avea una sau mai multe forme şi poate avea zero sau mai multe sensuri.

Clasa *Entry Relatio*n este o clasă de referințe încrucişate care poate lega două sau mai multe intrări lexicale din cadrul aceluiaşi lexicon sau din lexicoane diferite. Atributele conținute în această clasă descriu tipul relațiilor.

Clasa *Form* reprezintă o variantă lexicală a formei scrise sau vorbite a intrării lexicale. O formă conține un şir Unicode care reprezintă forma cuvântului şi categoriile de date care descriu atributele formei cuvântului. O clasă *Form* poate conține mai multe variante ortografice (lema, pronunția, despărțirea în silabe). Clasa *Form* are două subclase: *Lemmatised Form* (care poate conține numai leme) şi *Inflected Form* (care poate conține numai forme flexionate).

---

[14] http://tagmatica.fr/doc/ISO24613cdRev9.pdf




Dacă există mai multe forme ortografice de reprezentare a cuvântului (transliterații, romanizări, pronunții), clasa *Form* poate fi asociată cu clasa *Representation Frame*. Aceasta conține o reprezentare ortografică specifică şi una sau mai multe categorii de date care descriu atributele acelei reprezentări.

Clasa *Sense* conține atributele care descriu semnificația intrării lexicale. Ea permite reprezentarea ierarhică a sensurilor: o parte a sensului poate fi în relație cu o altă parte a aceluiaşi sens.

Clasa *Sense Relation* este o clasă de referințe încrucişate care poate lega două sau mai multe sensuri pentru o singură limbă sau între mai multe lexicoane. Ea poate conține atribute care descriu tipul relațiilor semantice.

**2.1.1.2. Extensiile LMF**

Toate extensiile se conformează pachetului central LMF. O extensie nu poate fi utilizată pentru a reprezenta datele lexicale independent de pachetul central. În funcție de tipul datelor lingvistice, o extensie poate depinde de o altă extensie. Din punctul de vedere al UML, o extensie este un pachet UML.

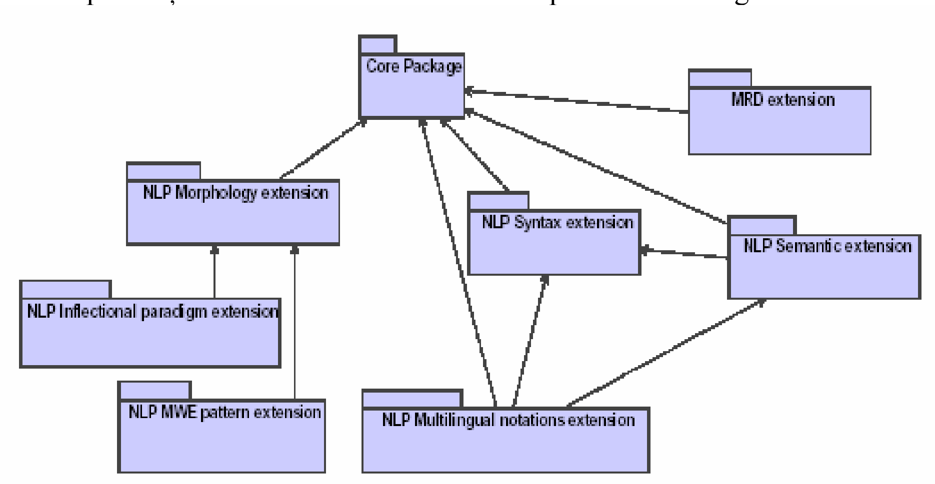
Mecanismul realizării unei extensii include:

– crearea subclaselor pe baza principiilor de modelare UML;
– adăugarea de noi clase;
– restricții asupra cardinalității şi tipului asocierilor;
– permiterea punctelor diferite de ancorare pentru asocieri;
– selecția categoriilor de date.

Extensiile NLP ale modelului LMF cuprind:

– extensia NLP pentru morfologie;
– extensia NLP pentru paradigma de flexiune;
– extensia NLP pentru modelul multi-cuvânt (MWE – MultiWord Expression);
– extensia NLP pentru sintaxă.

Dependențele între diferitele extensii sunt reprezentate în diagrama următoare[15]:




*2.1.2. LMF pentru datele terminologice*

Deşi LMF a fost conceput pentru reprezentarea informațiilor lexicale, există posibilitatea ca el să poată fi folosit şi pentru reprezentarea informațiilor terminologice, realizându-se astfel dezideratul de a avea o reprezentare uniformă a celor două tipuri de informații.

Deşi în G. Francopoulo *et alii* (2006a) se foloseşte expresia *LMF pentru lexicoane specializate*, caracteristicile prezentate sunt cele specifice colecțiilor de date terminologice:

1. un număr mare de expresii multi-cuvânt (MWE – Multiword Expression) – termenii sunt, în general, compuşi din mai multe cuvinte;
2. un număr mare de variante ortografice, care includ abrevierile şi acronimele – standardele terminologice prezintă posibilele forme prescurtate ale termenilor: forma scurtă, forma troncată, sigla, acronimul, abrevierea;

[15] http://tagmatica.fr/doc/ISO24613cdRev9.pdf

3. includerea informațiilor specifice domeniului: definițiile terminologice, coduri specifice – standardele terminologice prevăd informații de descriere noțională – definiții, contexte de utilizare etc.;
4. specificarea domeniului de cunoştințe – informație esențială în terminologii.

Soluțiile pe care le oferă LMF pentru tratarea caracteristicilor terminologice sunt:

a) expresiile multi-cuvânt pot fi folosite, în particular, pentru a reprezenta termeni;
b) diferitele forme prescurtate ale termenilor pot fi reprezentate ca variante ortografice sau cu ajutorul relației de sinonimie din LMF;
c) fiecare element LMF poate fi îmbogățit cu perechi atribut-valoare care să reprezinte informațiile de domeniu/subdomeniu, definiție etc.

### 2.2. Modelul OLIF – Open Lexicon Interchange Format

OLIF este un standard deschis pentru codificarea informațiilor lexicale şi terminologice. Conceput în anii '90 ca un mijloc de a asigura interschimbul de informații între aplicațiile NLP (în special traducerea automată) şi bazele de date terminologice, OLIF a evoluat într-un standard care oferă mijloace diverse pentru facilitarea reprezentării şi interschimburilor datelor lingvistice.

Prototipul OLIF a fost dezvoltat în cadrul proiectului OTELO (Open Translation Environment for Localization) al Uniunii Europene (încheiat în 2000). În martie 2000 s-a constituit consorțiul OLIF[16], al cărui scop este de a elabora un standard bazat pe prototipul OLIF.

Versiunea a doua a OLIF a fost elaborată în cadrul inițiativei SALT (Standards-based Access service to multilingual Lexicons and Terminologies); SALT este un consorțiu de instituții academice, guvernamentale, comerciale din Europa şi Statele Unite, care are ca scop elaborarea de formate şi instrumente care să faciliteze interschimbul de informații între bazele de date terminologice şi lexicoanele utilizate de aplicațiile NLP.

Implementat ca XML Schema (OLIF v.2.1) şi ca XML DTD (OLIF v.2), prototipul OLIF oferă un set reprezentativ de trăsături lexicale şi terminologice utile în aplicații ca:

- interschimbul de informații lexicale/terminologice;
- managementul lexicoanelor şi terminologiilor;
- extragerea termenilor [engl.: *term extraction*];
- limbaj controlat;
- regăsirea informației [engl.: *information retrieval*];
- dezvoltarea glosarelor;
- crearea de ontologii.

[16] Consorțiul **OLIF** este format din SAP, Microsoft, LionBridge X, Trados, Systran, IBM, DFK, BASIS Technology, SLS Smart Logic Solution.

*2.2.1. Structura generală a OLIF*

Datele sunt organizate în trei grupe esenţiale de informaţii:

1. Antetul (*header*) conţine informaţii comune tuturor intrărilor lexicale/terminologice;
2. Corpul (*body*) conţine intrările lexicale/terminologice propriu-zise;
3. Resursele partajate (*shared resources*) conţin informaţii suplimentare referite de intrările lexicale/terminologice (de exemplu, sursele bibliografice).

*2.2.2. Structura corpului (body) unui fişier OLIF*

Corpul fişierului OLIF este o listă de intrări care conţin date grupate conform caracterului lingvistic/lexical/terminologic al informaţiilor reprezentate. Deoarece scopul principal al OLIF este acela de a oferi o punte de legătură între lexicoanele utilizate în aplicaţiile de procesare a limbajului natural (NLP) şi aplicaţiile de management al terminologiilor, el a fost proiectat având în vedere atât punctul de vedere lexical, cât şi cel terminologic. Structura unei intrări (*entry*) OLIF reflectă, prin urmare, acest caracter hibrid, fără a fi orientată pe lemă, ca majoritatea lexicoanelor, şi fără a fi orientată pe concept, ca majoritatea bazelor de date terminologice. Pentru a reprezenta atât informaţiile necesare lexicoanelor, cât şi cele din terminologii, în OLIF se optează pentru o structură flexibilă, orientată pe semnificaţia cuvântului [engl.: *word-sense orientation*]. O intrare este o *colecţie de informaţii monolingve* asupra unei *semnificaţii specifice* a unui cuvânt sau a unei expresii; între intrări pot fi *referinţe încrucişate*, care reflectă diferite *relaţii* în cadrul aceleiaşi limbi; echivalenţele între intrările dintr-o limbă oarecare şi intrările corespunzătoare în altă limbă sunt date sub formă de *relaţii de transfer*.

O intrare OLIF are următoarea componenţă de informaţii:

- un grup unic de informaţii, a cărui prezenţă este *obligatorie*, care conţine datele corespunzătoare unei singure limbi, numit **monolingual (*mono*)**;
- o mulţime *opţională* de grupuri de informaţii – **cross-reference** (***cross-Refer***) – care reprezintă relaţiile intrării respective cu alte intrări corespunzătoare aceleiaşi limbi; fiecare grup **cross-reference** reprezintă o singură relaţie; pot exista mai multe grupuri cross-reference ale aceleiaşi intrări; în aceste grupuri sunt reprezentate relaţiile de sinonimie, omonimie, antonimie etc.;
- o mulţime *opţională* de grupuri de informaţii – **transfer** – care reprezintă relaţiile dintre intrarea respectivă şi intrările corespunzătoare din alte limbi; fiecare grup **transfer** reprezintă o singură relaţie de transfer, unidirecţională; în aceste grupuri sunt specificate legăturile spre echivalenţele în alte limbi corespunzătoare intrării respective;
- un grup *opţional* de categorii de date generale – **general** (***generalDC***) care poate fi inclus în oricare din grupurile **monolingual**, **cross-reference** sau **transfer**.

În cele ce urmează, prezentăm acele aspecte ale OLIF care ne interesează în demersul nostru de inventariere a trăsăturilor lexico-morfologice şi sintactice care să fie reprezentate în lexicon pentru realizarea analizei sintagmatice. Prin urmare, nu vom inventaria trăsăturile specifice reprezentării conceptelor şi termenilor. De asemenea, vom extrage din standardul OLIF doar partea care interesează reprezentarea monolingvă, ignorând aspectele care vizează transferul între mai multe limbi.

### *2.2.3. Categoriile esențiale de date*

Categoriile esențiale de date [engl.: *key data categories*] – **keyDC** – specifică semnificația unei intrări date într-o limbă anume şi sunt obligatorii în grupul **monolingual** al fiecărei intrări.

Există cinci categorii esențiale de date:

- *forma canonică* [engl. *canonical form*] – ***canForm*** – şirul de caractere corespunzător intrării, reprezentat în formă canonică, conform specificațiilor OLIF; este, în esență, forma lematizată a intrării;
- *limba* [engl. *language*] – ***language*** – este specificarea limbii corespunzătoare intrării;
- *partea de vorbire* [engl. *part of speech*] – ***ptOfSpeech*** – specifică partea de vorbire a intrării;
- *domeniul* [engl. *subject field*] – ***subjField*** – specifică domeniul de cunoştințe corespunzător intrării;
- *semnificația* [engl. semantic reading] – ***semReading*** – identificatorul clasei semantice, utilizat pentru a distinge semnificații diferite ale intrărilor care au valori identice pentru *forma canonică*, *limbă*, *parte de vorbire* şi *domeniu*.

$$\begin{bmatrix} \textbf{entry} & \begin{bmatrix} \textbf{mono} & \begin{bmatrix} \textbf{keyDC} & \begin{bmatrix} \text{canForm} & \textit{tabel} \\ \text{language} & \textit{ro} \\ \text{ptOfSpeech} & \textit{noun} \\ \text{subjField} & \textit{general} \\ \text{semReading} & 32 \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

*Exemplu*[17] *de categorii esențiale de date, dat sub formă de structuri de trăsături*

Versiunea 2.0/2.1 a modelului a fost extinsă cu scopul de a permite atât reprezentarea datelor orientată pe concept (ca în majoritatea colecțiilor de date terminologice), cât şi reprezentarea orientată pe leme (ca în colecțiile de date lexicografice). Identificatorii de lemă şi de concept sunt prezenți la nivelul superior

[17] Exemplele sunt adaptate după Susan McCormick, OLIF Consortium *The Structure and Content of the Body of an OLIF v.2.0/2.1 File*

al intrării (*entry*). Identificatorul de concept permite organizarea intrărilor ca semnificaţii echivalente asociate aceluiaşi concept, iar identificatorul de lemă permite organizarea intrărilor cu semnificaţii diferite, care au aceeaşi lemă:

$$\left[\textbf{entry}\quad \left[\begin{array}{l}\left\{\begin{array}{l}\text{conceptID}\\ \text{lemaID}\end{array}\right\}\\ \textbf{mono}\end{array}\right]\right]$$

Informaţiile relative la un singur cuvânt sau termen (care sunt, prin urmare, asociate unei singure limbi), sunt reprezentate în intrarea OLIF în grupul ***mono*** (**monolingual**); acest grup cuprinde datele esenţiale (***keyDC***), obligatorii, precum şi alte categorii de date (***monoDC***), grupate, la rândul lor, în funcţie de natura informaţiilor – lingvistică, lexicală sau terminologică – în date administrative (***monoAdmin***), morfologice (***monoMorph***), sintactice (***monoSyn***) şi semantice (***monoSem***). Datele esenţiale ***keyDC*** identifică în mod univoc intrarea.

Pentru identificarea intrărilor şi pentru simplificarea reprezentării referinţelor încrucişate şi a transferurilor, modelul OLIF prevede mai multe posibilităţi de a specifica identificatorii numerici (care pot fi referiţi în cadrul relaţiilor), asociaţi fie grupului **mono** (***monoUserId*** sau ***monoUniversalId***), fie categoriilor esenţiale de date (***keyDCUserId*** sau ***keyDCUniversalId***), fie intrării (***lemmaUserId***).

**Categoriile de date**

Categoriile de date din OLIF referitoare la informaţiile lexicale sunt ilustrate în figura următoare; prezentarea este sub forma structurilor de trăsături (categoriile de date scrise cu aldine sunt obligatorii), astfel: [denumire *valoare* (explicaţie)] .

Structura intrării are forma:

$$\left[\text{entry}\quad \left[\begin{array}{lll}
\text{lemmaUserId} & \textit{nb} & \left(\begin{array}{l}\text{identificator}\\ \text{al utilizatorului}\\ \text{pentru lemă}\end{array}\right)\\
\textbf{mono} & \left[\begin{array}{lll}
\text{monoUserId} & \textit{nb} & \left(\begin{array}{l}\text{identificator al}\\ \text{utilizatorului}\end{array}\right)\\
\text{monoUniversalId} & \textit{nb} & \left(\begin{array}{l}\text{identificator}\\ \text{universal}\end{array}\right)\\
\textbf{keyDC} & [...] & (\text{date esenţiale})\\
\text{monoDC} & [...] & \left(\begin{array}{l}\text{alte categorii}\\ \text{de date}\end{array}\right)\\
\text{generalDC} & [...] & (\text{informaţii generale})
\end{array}\right] & \left(\begin{array}{l}\text{grupează datele}\\ \text{monolingve}\\ \text{ale unei intrări}\end{array}\right)
\end{array}\right] \quad (\ )\right]$$

unde grupul monolingual – **mono** este:

| | | | |
|---|---|---|---|
| mono | monoUserId | *nb* | (identificator al utilizatorului) |
| | monoUniversalId | *nb* | (identificator universal) |
| | **keyDC** | [...] | (date esențiale) |
| | monoDC | [...] | (alte categorii de date) |
| | generalDC | [...] | (informații generale) |

Categoriile esențiale de date, *keyDC* au următoarea structură:

| | | | | | |
|---|---|---|---|---|---|
| **keyDC** | keyDCUserId | *nb* | | (identif. al utilizatorului) | ( ) |
| | keyDCUniversalId | *nb* | | (identif. universal) | |
| | **canForm** | *string* | | (forma canonică) | |
| | **language** | *string* | | (limba) | |
| | **ptOfSpeech** | ⟨ noun | (substantiv) | (partea de vorbire) | |
| | | verb | (verb) | | |
| | | adj | (adjectiv) | | |
| | | adv | (adverb) | | |
| | | prep | (prepoziție) | | |
| | | conj | (conjuncție) | | |
| | | det | (determinant) | | |
| | | part | (particulă verbală) | | |
| | | auxverb | (verb auxiliar) | | |
| | | pron | (pronume) | | |
| | | punc | (semn de punctuație) | | |
| | | other ⟩ | (alta) | | |
| | **subjField** | string | | (domeniu) | |
| | **semReading** | string | | (definiție) | |

Categoriile de date asociate grupului mono (*monoDC*) sunt date suplimentare, care pot fi: informații administrative – ***monoAdmin***, informații morfologice – ***monoMorph***, informații sintactice şi semantice – ***monoSyn*** şi, respectiv, ***monoSem***.

Astfel, informațiile administrative relative la datele lexicografice conțin date despre despărțirea în silabe – ***syllabification,*** despre structura şirului care formează intrarea – ***entryFormation*** (abreviere, acronim, cuvânt simplu, cuvânt compus, expresie, nespecificat), despre tipul expresiei – ***phraseType*** (expresie multi-cuvânt, expresie fixă, lexicalizată, colocație, idiom, nespecificată), autorul intrării – ***originator*** sau statutul intrării – ***adminStatus*** (intrare nouă, intrare verificată, implicită, exclusiv pentru MT, învechită, nespecificat).

*Informații morfologice* – ***monoMorph*** sunt cele care se referă la transcrierea structurii morfologice – ***morphStruct***, la clasa de flexiune – ***inflection***, cele care indică centrul sintagmei – ***head***, genul gramatical – ***gender***, cazul – ***case***, numărul – ***number***, persoana – ***person***, timpul – ***tense***, modul – ***mood***, aspect – ***aspect***, gradele de comparație – ***degree*** sau tip de verb auxiliar – ***auxType*** (în funcție de limbă).

*Informații sintactice* – ***monoSyn*** sunt cele care descriu comportarea generală a şirului de intrare – ***synType***, poziționarea nemarcată a şirului de intrare, din punct de vedere sintactic – ***synPosition***, descriu tranzitivitatea verbelor – ***transType***, structura de constituenți a unui şir multi-cuvânt – ***synStruct***, valența şirului de intrare – ***synFrame*** etc.

Informații semantice – ***monoSem*** se referă la definiția – ***definition***, genul – ***natGender*** (masculin, feminin, nespecificat) şi la o categorie care reprezintă tipul unei intrări care respectă o anumită clasificare semantică – ***semType***.

Grupul de date generale – *generalDC* are forma:

$$\begin{bmatrix} \text{generalDC} & \begin{bmatrix} \text{updater} & \textit{string} & \begin{pmatrix} \text{persoana care a} \\ \text{operat modif.} \end{pmatrix} \\ \text{modDate} & \textit{date} & (\text{data modif.}) \\ \text{example} & \textit{string} & (\text{exemplu}) \\ \text{usage} & \textit{string} & (\text{utilizare}) \\ \text{note} & \textit{string} & \begin{pmatrix} \text{comentariu al} \\ \text{lexicografului/} \\ \text{terminologului} \end{pmatrix} \end{bmatrix} & \begin{pmatrix} \text{categorii de} \\ \text{date generale} \end{pmatrix} \end{bmatrix}$$

## 3. Informațiile lexico-morfologice şi sintactice utilizate în proiectul SIASTRO

Proiectul **SIASTRO** continuă cercetările grupului *RoLingva* de la Software ITC SA asupra morfologiei. Acesta a realizat *Dicționarul Morfologic Român*[18],

[18] http://www.rolingva.ro/aplicatii_dictionar.php

dicționar în format electronic, care acoperă lexicul *DEX* şi care conține, printre alte componente, un analizor morfologic.

În ce priveşte modelele formale de reprezentare a informației lexicale s-a considerat că cel mai apropiat de structura existentă a dicționarului morfologic şi de cerințele impuse de analiza sintagmatică a limbii române este OLIF.

Intrările *Dicționarul Morfologic Român* au asociate atribute lexico-morfologice şi fonologice. O parte dintre aceste atribute pot servi scopului propus; altele (de exemplu, cele fonologice) nu sunt necesare analizei sintagmatice; altele trebuie rafinate. Pentru realizarea analizei sintagmatice, fiecare clasă lexico-gramaticală trebuie îmbogățită cu atribute sintactico-semantice care nu sunt prezente în *Dicționarul morfologic*.

În vederea stabilirii mărcilor specifice claselor şi subclaselor lexico-morfologice, care să facă posibilă realizarea automată a analizei sintagmatice, s-a procedat la examinarea atributelor relevante ale fiecărei clase şi a valorilor acestora.

Pentru selectarea/stabilirea atributelor şi valorilor ce privesc categoriile morfologice s-a utilizat în principal GALR 2005. S-a apelat, de asemenea, şi la materialul ştiințific oferit de D. D. Draşoveanu, G. G. Neamțu şi D. Bejan, ca şi la cadrul în sens larg şi fundamental teoretic oferit de lingvistica integrală a lui E. Coşeriu. Există cazuri în care s-a considerat că opțiunile terminologice şi/sau teoretice ale acestora sunt mai adecvate obiectivelor proiectului decât cele din GALR 2005, iar acestea sunt menționate explicit în studiul dedicat categoriilor morfologice în cauză.

Deoarece ca primă aplicație a analizorului sintagmatic este prevăzută realizarea unui extractor de termeni, ne-am orientat spre studierea acelor formate de reprezentare a datelor lexico-morfologice şi sintactice care să permită şi extinderea spre bazele de date terminologice. În ce priveşte terminologia, ca teorie, am adoptat concepția noțională (orientată pe concept), care stă la baza modelelor formale analizate.

Premisele elaborării tehnologiei lingvistice pentru limba română au fost create de studiile teoretice, realizate de grupul *RoLingva* asupra morfologiei. A fost realizat primul model formalizat al morfologiei cu reguli şi metode care pot acoperi în totalitate sistemul flexionar din limba română. În afară de informațiile referitoare la elementele morfologice şi ortografice, modelul păstrează informații incipiente despre elemente fonetice, cum ar fi silabația fonetică, accent, împreună cu regulile de migrare a acestuia în cadrul flexiunii. Modelul morfologic respectă regulile gramaticale clasice şi a fost realizat ca un sistem deschis pentru a putea acoperi cu uşurință şi alte domenii ale gramaticii limbii române. El permite extinderile necesare pentru a prelua informațiile care vor rezulta din formalizarea sintaxei limbii române.

Limba română este o limbă care flexionează puternic şi permite realizarea de numeroase cuvinte noi prin sufixare şi prefixare, cuvinte care, la rândul lor, pot avea propria flexiune. Numărul formelor flexionate este mare şi, practic, se poate pierde controlul asupra acestora, dacă se va face un simplu inventar al tuturor formelor însoțite de categoria gramaticală reprezentativă. Din acest motiv, în

tezaurul de cunoştinţe lingvistice s-a preferat obţinerea acestor forme prin metode procedurale, pe bază de reguli, plecând de la o *rădăcină* condensată, liste de *sufixe* şi *prefixe*, şi o *clasă flexionară părinte*, care pot descrie în totalitate o intrare lexicală. Combinaţiile care se pot realiza astfel sunt foarte numeroase şi pot acoperi un lexic foarte bogat.

Pentru proiectul **SIASTRO** s-a făcut o documentare asupra atributelor existente în tezaurul de cunoştinţe lingvistice, ţinându-se cont de importanţa lor în lexiconul care se va crea. De asemenea, s-a făcut o paralelă între atributele utilizate de sistemele standardizate şi atributele care pot fi preluate din tezaurul de cunoştinţe lingvistice.

S-a creat, în felul acesta, un model de reprezentare a atributelor morfo-sintactice şi s-au dezvoltat procedurile care generează lexiconul îmbogăţit cu aceste atribute.

Astfel, pornind de la modelul OLIF, s-a creat o **structură de intrare** de forma:

| | | |
|---|---|---|
| intrare | IdentificatorLema | |
| | **Atribute** | NUMAR_DE_IDENTIFICARE |
| | | **ATRIBUTE_CHEIE** |
| | | ATRIBUTE_MORFO_SINTACTICE |
| | | INFORMATII_ADMINISTRATIVE |

*Structura intrării SIASTRO*

în care **atributele cheie** sunt:

| | | | |
|---|---|---|---|
| **ATRIBUTE_CHEIE** | INDEX_LEMA | | |
| | **LEMA** | *string* | |
| | **PARTE_VORBIRE** | sbt | (substantiv) |
| | | adj | (adjectiv) |
| | | vrb | (verb) |
| | | adv | (adverb) |
| | | cnj | (conjuncţie) |
| | | prep | (prepoziţie) |
| | | num | (numeral) |
| | | pron | (pronume) |
| | | adjpr | (adjectiv pronominal) |
| | | art | (articol) |
| | | inj | (interjecţie) |
| | | mrf | (morfem cazual) |
| | **DOMENIU** | string | |
| | **GLOSA** | string | |

*Atribute-cheie SIASTRO*

Pentru fiecare clasă lexico-gramaticală s-au stabilit atribute şi valori specifice. Aceste atribute se împart, la rândul lor, în mai multe categorii:

- atribute lexico-morfologice propriu-zise (de exemplu, TIP, TEMA pentru clasele lexico-gramaticale flexionale, GEN, pentru substantive etc.);
- atribute sintactice: VALENȚA, POZIȚIE (pe care o poate avea adjectivul în raport cu substantivul determinat), DETERM (reprezintă comportamentul determinării adjectivului antepus), PROXIMITATE_REGENT (se aplică la adjectivele pronominale care pot fi postpuse şi reprezintă proximitatea față de regent) etc.;
- atribute care facilitează procedura de generare a lexiconului din informațiile existente în *Dicționarul Morfologic*: CONVERSIE (dacă este posibilă conversia de la o clasă lexico-gramaticală la alta, cu specificarea claselor respective), PREFIXE (dacă admite derivarea cu prefixe, iar în caz afirmativ, se specifică prefixele respective), MEMBRU ÎN LOCUȚIUNE şi STRUCTURA LOCUȚIUNII etc.

## 4. Concluzii

În vederea stabilirii structurii lexiconului SIASTRO, studiile s-au desfăşurat pe două direcții: una de analiză a modelelor formale existente pentru descrierea datelor lexicale şi terminologice, iar a doua, de analiză aprofundată a atributelor lexicale, morfologice şi sintactice necesare realizării analizei sintagmatice pentru textele scrise în limba română.

Până în prezent s-au realizat procedurile de înscriere a atributelor noi în dicționar şi a celor de generare a lexiconului necesar analizorului sintagmatic.

În viitor urmează să se adauge în lexicon structurile multi-cuvânt şi informațiile de descriere terminologică. Aşa cum este concepută acum structura lexiconului, aceste extensii se vor integra în mod natural în lexiconul existent.

## REFERINȚE BIBLIOGRAFICE


ISLE Meta Data Initiative 2003 = *Metadata Elements for Lexical Descriptions*, Version 1.1c, MPI Nijmegen, http://www.mpi.nl/IMDI/documents/Proposals/IMDI_Lexicon_1.1c.pdf

ISO/CD 24613:2006 *Language resource management – Lexical markup framework(LMF)*. www.tc37sc4.org/new_doc/ISO_TC37_N130_rev9_LMF_15March2006.pdf, http://tagmatica.fr/doc/ISO24613cdRev9.pdf

Antoni-Lay, Marie-Hélène, Gil Francopoulo, Laurence Zaysser (1994), *A Generic Model For Reusable Lexicons: The Genelex Project*. http://perso.orange.fr/laurence.zaysser/llc94.html

Calzolari, Nicoletta, Alessandro Lenci, Francesca Bertagna, Antonio Zampolli (2002), *Broadening the Scope of the* **EAGLES/ISLE** *Lexical Standardization Initiative*, in *Proceedings of the 3rd*

*workshop on Asian language resources and international standardization,* Volume 12. http://ucrel.lancs.ac.uk/acl/W/W02/W02-1204.pdf

Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria (2006a) *LMF for Multilingual, Specialized Lexicons*, in Pierre Zweigenbaum, Stefan Schulz and Patrick Ruch (editors), *LREC 2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine*, Genova, Italy, 2006, ELDA. http://estime.spim.jussieu.fr/~pz/lrec2006/Francopoulo.pdf

Francopoulo, Gil, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, Claudia Soria (2006b) *Lexical Markup Framework (Lmf) For Nlp Multilingual Resources*, **i**n *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 1–8, Sydney, July 2006. © 2006 Association for Computational Linguistics

Jacquemin, Christian, Evelyne Tzoukermann (1999), *NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax*, in T. Strzalkowski (ed), *Natural Language Processing Information Retrieval,* p. 25-74, Kluwer, Boston, MA, 1999 http://citeseer.ist.psu.edu/cache/papers/cs/22948/http:zSzzSzwww.limsi.frzSzIndividuzSzjacquemizSzFTPzSzjacqtzou-NLIR97.pdf/jacquemin99nlp.pdf

Lieske, Christian (2001), *The Open Lexicon Interchange Format (OLIF) Comes of Age*, Machine Translation Summit VIII 2001, www.olif.net/documents/olifMtSummitVIII.pdf

Mc Cormik, Susan, Christian Lieske, Alexander Culum (2004), *OLIF v.2: A Flexible Language Data Standard* http://www.olif.net/documents/olifMtSummitVIII.pdf

Mc Cormik, Susan (2005), *The Structure and Content of the Body of an OLIF v.2.0/2.1 File*. www.olif.net

Melby, Alan K. (2000), *SALT: Standards-based Access service to multilingual Lexicons and Terminologies*. www.ttt.org

Monachini, Monica, Francesca Bertagna, Nicoletta Calzolari, Nancy Underwood, Costanza Navarretta (2003), *Towards a Standard for the Creation of Lexica* www.elra.info/services/standard_lexica.pdf

OLIF Consortium (2005), **OLIF** – *The Open XML Language Data Standard* www.olif.net

Peters, Wim (2002), *Resurse Lexicale*. http://phobos.cs.unibuc.ro/roric/Ro/lex_introduction.html

Romary, Laurent (2001), *Un modèle abstrait pour la représentation de terminologies multilingues informatisées: TMF – Terminological Mark-up Framework*, in *Cahiers GUTenberg* no. 39-40, mai 2001, http://www.gutenberg.eu.org/pub/GUTenberg/publicationPDF/39-romary.pdf

Romary, Laurent, Marc Van Campenhoudt (2001), *Normalisation des échanges de données en terminologie: le cas des relations dites «conceptuelles»*, in *Conférence TIA-2001*, Nancy, 3 et 4 mai 2001, www.termisti.refer.org/tia4.pdf

Romary, Laurent, Salmon-Alt S., Francopoulo G. (2004), *Standards going concrete: from LMF to Morphalou*, in *Workshop on Electronic Dictionaries*, Coling 2004, Geneva, Switzerland, http://acl.ldc.upenn.edu/coling2004/W10/pdf/4.pdf

Segura Bedmar et alii 2006 = Segura Bedmar, Isabel, José L. Martínez Fernández, Paloma Martínez, *Including deeper semantic information in the Lexical Markup Framework: a proposal*, in *Proceedings of 5th Slovenian and 1st international Language Technologies Conference*, 2006 ***IS-LTC'06*** nl.ijs.si/is-ltc06/proc/15_Segura.pdf

Tămâianu-Morita, Emma (2008) *Tipologia sintagmelor în modelul D. D. Draşoveanu. Posibile aplicaţii în proiectul SIASTRO, supra*, p. 137-150

TEI *Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition*, Edited by C. M. Sperberg-McQueen and Lou Burnard; XML conversion by Syd Bauman, Lou Burnard, Steven DeRose, and Sebastian Rahtz © 2001, 2002, 2004 by the TEI Consortium www.tei.org

Wright, Sue Ellen (1999) *TO11: SALTing the Alphabet Soup*. ***TC Forum 1998-2001*** *- http://www.tc-forum.org - file last updated 17 Oct 1999*

## FORMAL MODELS FOR REPRESENTATION OF LEXICAL AND TERMINOLOGICAL INFORMATION IN *SIASTRO* PROJECT (*Abstract*)

This paper outlines the research carried out within the SIASTRO project on the formal models for a uniform representation of lexical and terminological information. In this framework, we describe the structure of the lexicon that will be used in the phrase analysis of texts written in Romanian.

*Universitatea „Babeş-Bolyai”*
*Facultatea de Litere*
*Cluj-Napoca, str. Horea, 31*